# **Cricket Data Analysis**

**Objective:**

Analyse the batting performance of the players across various formats and perform cluster analyses based on features like Runs scored, batting average, strike rate etc…

**Data Cleaning and Transformation**

* Country and Players personal data is joined with the player statistics. Data for One Day International and T-20I formats are considered separately since the playing style is different considering aggression, strategy and tactics.
* Pandas Profiling: Data Profiling done using ydata\_profiling library. ProfileReport module is used to create a report based on various types of columns, records counts and values.
* Data is filtered for null values and infinite values.
* Innings cut-off is set, to filter the data. As a standard, only players with 20 or more innings played are selected to conduct the analysis. This helps to keep data quality and produce meaningful analyses
* Transform data to create new columns for the analyses.

**New Columns created**

* batting\_score: Harmonic mean of Strike rate and batting average score.
* years\_played: Number of years played by the player.
* Runs\_per\_year: Average runs scored by the player in a year.
* Runs\_per\_inning: Average runs scored by the player in an inning (not considering number of times player got out).
* Runs\_by\_6s: Runs scored by player only using hitting 6s.
* 6s\_runs\_percentage: Percentage of runs scored by player only using hitting 6s.
* Runs\_by\_4s: Runs scored by player only using hitting 4s.
* 4s\_runs\_percentage: Percentage of runs scored by player only using hitting 4s.

**Players Statistics Analysis** (Refer Jupyter Notebook file)

Analysis is performed on numerical values like runs, strike rate and average score. The catrgorical columns used for the same are mainly batting\_style, playing\_role and country.

**Outlier Analysis:**

A box plot is used to see the distribution of runs scored by players of left-handed and right-handed batting styles. While in T20I the distribution of runs was centred around 500 to 1000 runs, in ODIs it was in a range from 1500 to 2000. This is due to the longer playing time and higher number of players in ODI format. The plot showed many outliers mainly in right-handed batsmen. The outstanding players seem to outperform an average batsmen by miles according to this plot.

**Z-Score Analysis:**

A z-score model is used to find the variability of runs on both extremes. The upper boundary and lower boundary for both ODI and T20I formats are calculated.

|  |  |  |
| --- | --- | --- |
| Format | Upper Boundary | Lower Boundary |
| T-20I | 2743.625 | 5355.875 |
| ODI | 1801.375 | 3430.125 |

While for T20I format, 7 players scored above the upper boundary, it was 15 for ODI format. In case of lower boundary, there are around 130 players scoring below lower boundary value. This shows a left skew. This is because of the exceptional performance of some players who tends to be in the top tier. It could also be because other players are not able to score runs consistently.

**Runs Distribution:**

Histograms are plotted to see how players of various categories are scoring in both formats. Players from India, England, Australia, New Zealand and Bangladesh scored more than 30,000 runs in ODIs. These same teams lead the scoring in case of T20Is also. This is not necessarily a metric of better players. This can be because of absence of some players from other countries. This could also be due to the omission of some players in the innings cut-off filtering.

Clustered categorical plots are used to show runs scored by players of different countries split by the batting style. Most countries had right-handers leading the pack with more than 60% of the runs amassed in the total runs. Only instance of left-handed players scoring more than right-handers was seen for South Africa for the T20I format.

A Similar analysis was done with players of different playing roles in the batting order in place of countries. While in ODIs top order batsmen, wicket-keeper batsmen and allrounders scored more runs, T20I showed a higher scoring from opening batsmen too. A highly aggressive scoring pattern in T20I can be inferred from this. In ODIs the middle order batsmen had more runs scored by left handers. In the case of T20Is left-handed middle order and opening batsmen showed higher runs scored as opposed to right handers.

**Batting Averages and Strike Rates**

A distribution of player frequency across batting average and strike rates bins is plotted to see the range of scoring rates. Since T20Is are short matches with high intensity, power hitting character, a good strike rate will be in the range of 120 to 140. In the case of ODIs this will be in 65 to 85 range. The histogram shows a similar distribution of players with strike rates in the corresponding ranges. Batting average points to a more consistent player which is more important in longer formats as opposed to short format. A batting average in the range of 30 to 40 in ODIs and 20 to 35 in T20I is counted as normal. But there and exceptional players who score at an average greater than 50 in both formats. Since both strike rate and Batting average are leveraged different in the different formats, a new metric is created by taking a harmonic mean of these to columns. We have called it ***batting\_score***. This score helps to normalise the effect of both strike rate and batting average. From the pattern seen in the plots higher number of players fall in a range of 30 to 50 for T20Is and 40 to 60 for ODIs. Also see the density plots to see the distribution of these metrics for different playing roles.

**Career Duration**

When a histogram is plotted to see the average duration of a player’s career for different countries, a minimum duration of 5 years can be seen. To be fair these can be said only for the consistent players. While players from New Zealand and England play for and average of 10 years, West Indies cricketer plays less than 6 years in ODIs. In T20Is Zimbabwe players have highest average career duration with 9.1 years and Sri Lanka has the least with 5.8 years on average

**Clustering Analysis**

A cluster analysis is done for various metrics in the dataset to classify the performance of all the players across different formats. One of the most popular methods is K-means clustering. K means clustering calculates the distance for each point to the mean which is considered a centre point or centroid. The distance is then used to map each value to the cluster corresponding to the centroid.

To find the number of clusters, we use Elbow Method, which iterates through a range of values from which we plot the cluster with inertia which is the sum of the squared distance between each point and the centroid in a cluster. From this we select a cluster at the bend of the plot which we call the elbow. The cluster at elbow point is the number of clusters we use for plotting.

From the model, we calculate the centroid value for each column and plot it for different data point. This will form a cluster when plotted on a scatter plot. The evaluation of the clusters can be done using a Silhouette Score. It helps to see how effectively the clusters have been classified.

* Strike Rate: The clustering of Strike rate to Runs scored showed a highly compact distribution in clusters with lower strike rate. The distribution becomes sparser when the strike rate is increased. This denotes that the low scoring players also have a low strike rate. The high scoring players have a cluster with higher variance considering the difference in scores when you go up the scoring chart. This trend was similar for both formats.
* Batting Average: The Batting average vs runs scored scatter plots show a similar trend of compact classification in low scoring players. But the definition is higher in low the case of batting average because the strike rates are heavily diverse based on how aggressive the player is.
* Batter’s Metric: The *batting\_score* derivedmetric is closer to the batting average pattern in both T20 and ODI formats. The high average- high strike rate players in the pack tend to stray of into wide territories when potted on the scatter plots

**Evaluation**

The evaluation is done using the Silhouette score. The score may range from -1 to 1. A positive value shows appropriate association with the clusters. The Silhouette Score for each cluster is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| Format | Cluster 1 | Cluster 2 | Cluster 3 |
| ODI | 0.25 | 0.23 | 0.22 |
| T20I | 0.19 | 0.17 | 0.27 |

Overall Silhouette Scores:

T20I: 0.22

ODI: 0.23

The evaluation shows that the clusters are classified efficiently. Although we will be able to perform the classification more robustly if we have more data points and other features to support the data points. Some of them may include, runs scored while chasing a score or batting first. Percentage of runs scored by player out of the total runs scored by the team. A score based on the performance against strong and weak teams also can help with better analysis